Phylogeny of Viruses☆

Alexander E Gorbalenya, Leiden University Medical Center, Leiden, The Netherlands
C Lauber, Technische Universität Dresden, Dresden, Germany

Introduction: Evolution, Phylogeny, and Viruses 1
Tree Definitions 2
Phylogenetic Analysis 3
Applications of Phylogeny in Virology 6
Acknowledgements 7

Introduction: Evolution, Phylogeny, and Viruses

Biological species, including viruses, change through generations and over time in the process known as evolution. These changes are first fixed in the genome of successful individuals that give rise to genetic lineages. Due to either limited fidelity of the replication apparatus copying the genome or physico-chemical activity of the environment, nucleotides may be changed, inserted, or deleted. Genomes of other origin may also be a source of innovation for a genome through the use of specially evolved mechanisms of genetic exchange (recombination). Accepted changes, known as mutations, may be neutral, advantageous, or deleterious, and depending on the population size and environment, the mutant lineage may proliferate or go extinct. Overall, advantageous mutations and large population size increase the chances for a lineage to succeed. The fitness of a lineage is constantly re-assessed in the ever-changing environment and lineages that, due to mutation, became a success in the past could be unfit in the new environment. Due to the growing number of mutations accumulating in the genomes, lineages diverge over time, although occasionally, due to stochastic reasons or under similar selection pressure, they may converge.

The relationship between biological lineages related by common descent is called phylogeny; the same term also embodies the methodology of reconstructing these relationships. Phylogeny deals with past events and, therefore, it is reconstructed by quantification of differences accumulated between lineages. Due to the lack of fossils and (relatively) high mutation rates, viruses were not considered to provide a recoverable part of phylogeny until the advent of molecular data proved otherwise. Comparison of nucleotide and amino acid sequences, and, occasionally, other quantitative characteristics such as distances between three-dimensional structures of biopolymers, have been used to reconstruct virus phylogenies. Results of phylogenetic analysis are typically depicted in the form of a tree that may be used as a synonym for phylogeny. For instance, the all-inclusive phylogeny of cellular species is known as the Tree of Life (ToL) (Fig. 1A). More recently, two techniques, networks and forests of trees, are used to depict the complexity of phylogenetic relations and the uncertainty of phylogenetic inference, respectively.

With few exceptions, virus phylogeny follows the theory and practice developed for phylogeny of cellular life forms. For inferring phylogeny, differences between the sequences of species members, assumed to be of a discernable common origin, are analyzed. If species in all lineages evolve at a uniform constant rate, like clocks tick, their evolution conforms to a molecular clock model. The utility of this model in relation to viruses may be very limited. Rather, related virus lineages may evolve at different and fluctuating rates and some sites may mutate repeatedly, including reverse substitutions. As a result, reconstruction of a full record of change at all sites is associated with ever increasing uncertainty with each new mutation. Furthermore, the accumulation of inter-species residue differences may progress nonlinearly with the time elapsed. At present, our understanding of these parameters of virus evolution is poor and this limits our ability to assess the fit between a reconstructed phylogeny and the true phylogeny, with the latter practically remaining unknown for most virus isolates. This gap in our knowledge does not undermine the conceptual strength and utility of phylogenetic analysis for reconstructing the relationships between biological species including viruses.

The ultimate goal of virus phylogeny is reconstructing the relationships between 'all' virus isolates and species. For instance, cellular species form three (or two) compact domains (kingdoms) and their origin can be traced back to a common ancestor in the ToL, using either ribosomal RNA or a common set of single-copy genes. Such inference is not feasible for viruses due to their diversity and the lack of a universal molecular denominator (trait). Thus, reconstructing the comprehensive virus phylogeny may require comparisons that involve genomes of viral and cellular origins. This formidable task remains largely 'work in progress'. In fact, most efforts in virus phylogeny are invested in reconstructing the relationships at the micro, rather than grand, scale and they focus on well-sampled lineages that have practical (e.g., medical) relevance. Most recently, due to the advent of high-throughput next generation sequencing (NGS) and metagenomics, phylogeny of distant relations to characterize diverse viromes and the entire Virosphere has become an active area of research. Phylogeny itself or in combination with other data may provide a deep insight into virus evolution and diverse aspects of virus life cycles, including virus interactions with their hosts.

☆*Change History:* December 2016. AE Gorbalenya and C Lauber updated sections Introduction, Tree Definitions, Phylogenetic Analysis and Applications of Phylogeny in Virology, the reference section. Introduced Figures 1-4.

http://dx.doi.org/10.1016/B978-0-12-801238-3.95723-4

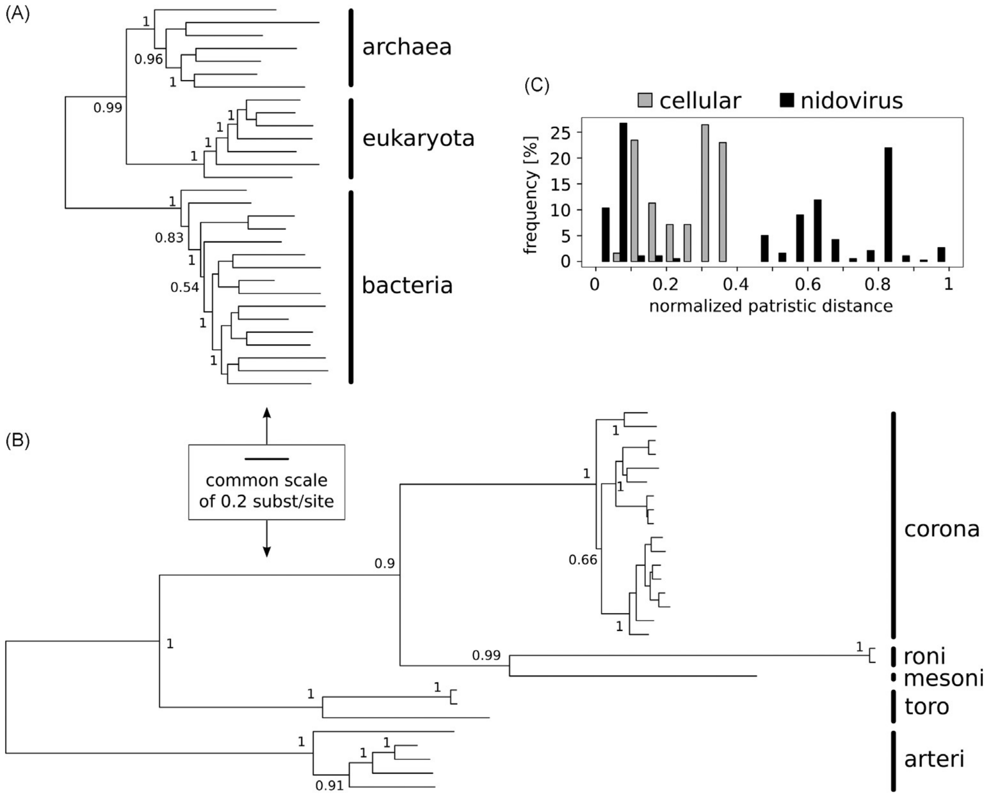

Fig. 1 Phylogeny of nidoviruses in comparison to the Tree of life (ToL). Bayesian phylogenies under the WAG amino acid substitution model with rate heterogeneity across sites and relaxed molecular clock with log-normal distribution of nidoviruses (A) and ToL (B) are drawn to a common scale of 0.2 amino acid substitutions per position. Major lineages are indicated by vertical bars and names; arteri: *Arteriviridae*, mesoni: *Mesoniviridae*, roni: *Roniviridae*, toro: *Torovirinae*, corona: *Coronavirinae*. Support values at basal internal nodes are posterior probability support values. (C) Distributions of pair-wise patristic distances extracted from (A) and (B). The combined set of distances was normalized relative to the largest distance that was set to one. Figure adapted from Lauber et al. (2013).

Our knowledge about contemporary virus diversity has been steadily advancing with new viruses being constantly described by systematic efforts as well as occasional discoveries. These developments indicate that only a small part of virus diversity has so far been unraveled and has become available for phylogenetic studies. It is also likely that many more lineages existed in the past; some of these lineages are likely to have ancestral relationships with contemporary lineages.

Tree Definitions

Species share similarity that varies depending on the rate of evolution and time of divergence. The entire process of generating contemporary species diversity from a common ancestor is believed to proceed through a chain of intermediate ancestors specific for different subsets of the analyzed species (Fig. 2). Typically, these ancestral sequences are estimated internally during the tree building process or are not required at all, depending on the method used. The relationship between the common ancestor, intermediate ancestors, and contemporary species may be likened to the relationship between, respectively, root, internal nodes, and terminal nodes (leaves) of a tree, an abstraction that is widely used for the visualization of this relationship (Fig. 2). Alignment of the contemporary sequences with the reconstructed tree side by side, like shown for the toy example in Fig. 2, may reveal the full chain of sequence changes that have happened during evolution which, however, is rarely the case for real data sets due to repeated substitutions and incomplete species sampling. Trees are also part of graph theory, a branch of mathematics, whose apparatus is used in phylogeny. Formally and due to a strong link between phylogeny and taxonomy, leaves may be called operational taxonomy units (OTUs) and internal nodes and roots, since they have not been directly observed, are known as hypothetical taxonomy units (HTUs). Nodes are connected by branches or edges.

The tree may be characterized by topology, length of branches, shape, and the position of the root (Fig. 3). The topology is determined by relative positions of internal and terminal nodes; it defines branching events leading to contemporary species diversity. If two or more trees obtained for different data sets feature a common topology, these trees are called congruent. The branch length of a tree may define either the amount of change fixed or the time passed between two nodes connected in a tree, and is known as 'additive' or 'ultrametric', respectively (Fig. 3B and C). The tree shape may be linked to particulars of the evolutionary process and reflect changes in population size and diversity due to genetic drift and natural selection. The position

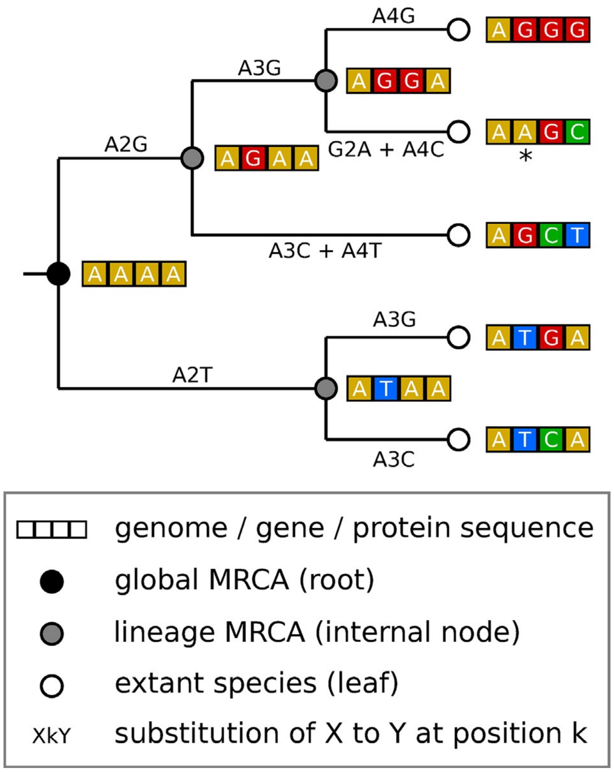

Fig. 2 Phylogenetic tree and molecular evolution. Shown is a toy example of evolution of an ancestral sequence (to the right of the black-filled circle) of length four into five extant sequences (to the right of the open circles) and the corresponding phylogenetic tree. The respective substitutions and sequence positions are indicated at the tree branches. Sequences at internal nodes (gray-filled circles) are reconstructions of the tree building process. Note the multiple substitutions at sequence position two in the second extant sequence resulting in reversion to the ancestral state (denoted with *).

of the root at the tree defines the direction of evolution. Species that descend from an internal node in a rooted tree form a lineage (cluster) and the node is called most recent common ancestor (MRCA) of the lineage that thus has a monophyletic origin (Fig. 2). The branch lengths and the root position may be left undefined for a tree that is then called 'cladogram' and 'unrooted tree', respectively (Fig. 3A).

Phylogenetic Analysis

Multiple alignments of polynucleotide or amino acid sequences representing analyzed species and maximized for similarity are traditionally used as input for phylogenetic analysis. The quality of alignment is among the most significant factors affecting the quality of phylogenetic inference. Due to the redundancy of the genetic code, changes in polynucleotide sequences are accumulated at a higher rate than those in amino acid sequences. In viruses, including RNA viruses, this difference is not counterbalanced by other local or global constraints on variation of genomes that are linked to e.g. di-nucleotide frequency or RNA secondary (tertiary) structure. Because of these differences, polynucleotide sequences are commonly used for phylogeny reconstruction of only those species that are closely related, while protein sequences, preserving better phylogenetic signal, may be used to infer phylogeny of distantly related species.

Differences between species, as calculated from alignment, may be quantified as either pairwise distances forming a distance matrix or position-specific substitution columns (discrete characters of states of alignment), the latter preserving the knowledge about location of differences. The respective methods dealing with these quantitative characteristics are known as distance and discrete (character state). The distance methods are praised for their speed and are considered a technique of choice for analysis of very large data sets, although character state methods caught up in this respect due to recent algorithmic advancements (see also below). Distance methods are often designed to converge on a unique phylogeny by clustering, with none others being even considered. The unweighted pair group method with arithmetic means (UPGMA) in which a constantly recalculated distance matrix is used to define the hierarchy of similarities through systematic and stepwise merging of most similar pairs at a time was the first technique introduced for clustering. The neighbor-joining (NJ) method uses a more sophisticated algorithm of clustering that minimizes branch lengths, and is the most popular among distance methods.

Although different trees may be compared in how they fit a distance matrix, it is character-based methods that are routinely used to assess numerous alternative phylogenies in search for the best one in a computationally intensive process. Due to the calculation time involved, assessing all possible phylogenies is found to be impractical for data sets including more than 10 sequences;

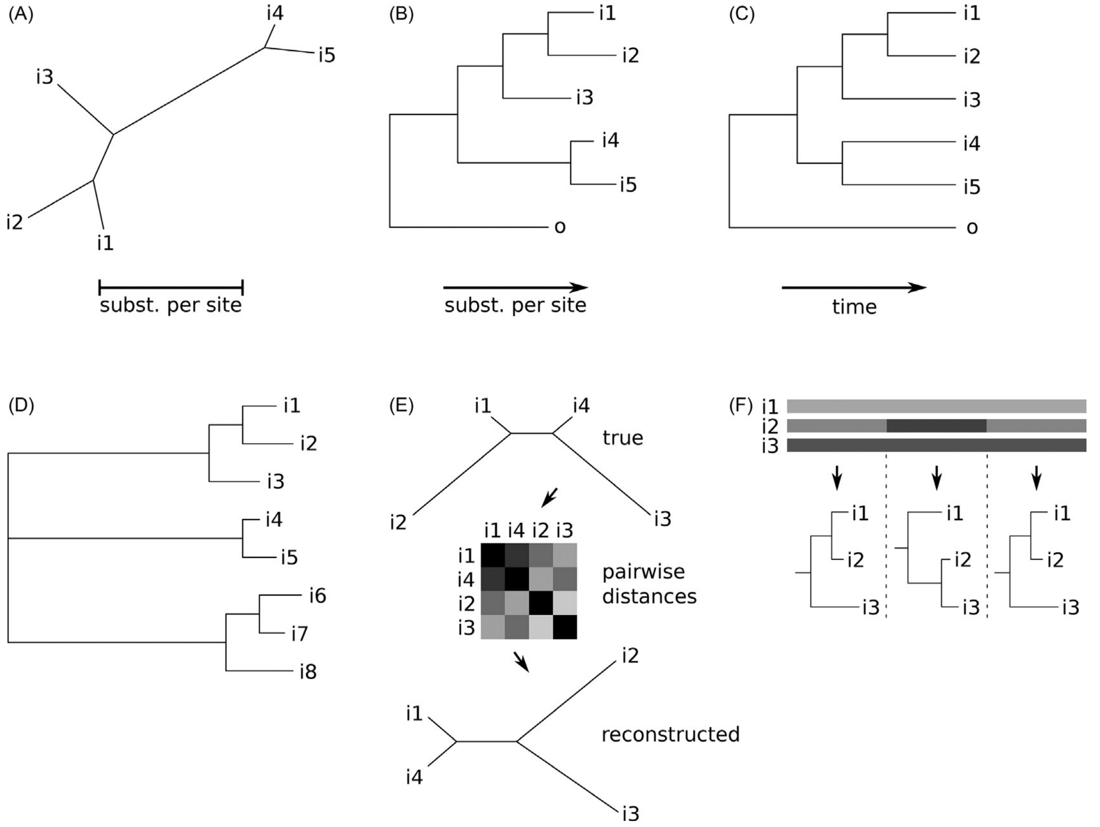

Fig. 3 Tree types and pitfalls of phylogeny reconstruction. (A) Unrooted tree for five hypothetical viruses that was reconstructed based on their gene or protein sequences. Branch lengths represent the amount of genetic change between two viruses typically measured in units of substitutions per site. The direction of evolution is undetermined. (B) The tree for the five ingroup viruses in A was rooted using an outgroup (o). The direction of evolution is from left to right. (C) The same tree as in B but with branches calibrated to represent time. Note that time does not necessarily correlate with the amount of genetic change (for instance, compare length of the branch leading to the cluster joining i4 and i5 with that of the same branch in B. (D) The relative positions of three highly divergent lineages is unresolved by the phylogeny (polytomy). (E) The true relationship of four hypothetical sequences (top) is not recovered by the phylogenetic reconstruction (bottom) due to long branch attraction involving i2 and i3. (F) Phylogenetic trees (bottom) reconstructed for three adjacent genomic regions (top) are different with respect to the position of i2 which was subject to a recombination event in middle genome region.

for larger data sets different heuristic approximations are used that may not guarantee a recovered phylogeny to be the best overall. There are two major criteria for selecting the best phylogeny using character-state based information through either maximum parsimony (MP) or maximum likelihood (ML). In MP analysis, a phylogeny with a minimal number of substitutions separating the analyzed species is sought. The ML analysis offers a statistical framework for comparing the likelihood of fitting different trees to the data under competing models of evolution with parameters including population size change and rate of mutation in search for one with the best fit. The latter approach is mathematically robust and its statistical power may also be used in combination with other techniques of tree generation. Recently, a Bayesian variant of the ML approach has gained popularity. It can utilize prior knowledge about the evolutionary process, like known substitution rates or clustering of species subsets or dates of species isolation, in combination with repeated sampling from subsequently derived hypotheses. The result of a Bayesian analysis is thus a forest of trees that reflects the uncertainty associated with the reconstructed phylogeny and which forms the basis to derive a consensus tree and statistic support for its branches. In phylogenetic analysis of viruses the dates of species isolation are often used to date the MRCA of the analyzed viruses under a Bayesian framework, while fossil information is routinely used to time-calibrate trees of cellular organisms. Bayesian methods have the highest computational cost due to their sampling approach and thus show the lowest speed, while realization of the similarly advanced ML algorithm may be largely comparable in speed to distance methods, allowing for the phylogenetic analysis of very large data sets like genome-wide tree reconstructions of cellular organisms or thousands of viruses.

One should keep in mind that different methods for phylogeny reconstruction can produce different trees, concerning both topology and branch lengths, for the same data set, although better agreement between ML and Bayesian trees is common, especially in respect to branch lengths (Fig. 4). None of the methods is considered superior to the other methods with respect to all aspects of phylogeny reconstruction, and which method to use under what circumstances is often a point of debate. A valid approach to gain further confidence in phylogenetic results is to apply several methods on the data and to only trust HTUs that are inferred by more than one method.

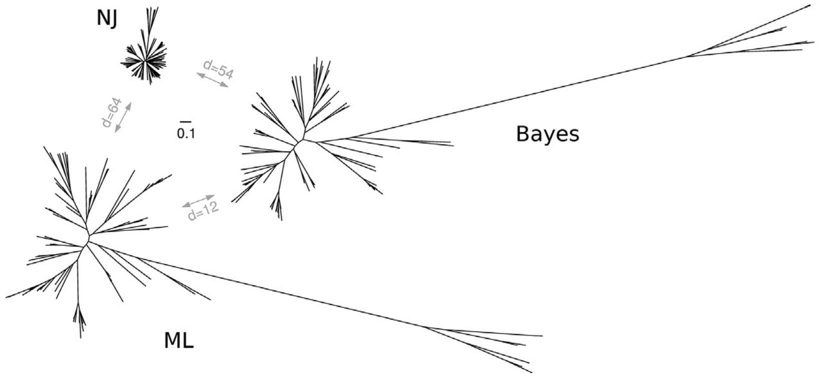

Fig. 4 Comparison of phylogenetic results between methods. Shown are unrooted NJ, ML and Bayesian trees reconstructed for the same dataset of 287 aligned VP1 protein positions of 93 polyomaviruses. The LG amino acid substitution model with site heterogeneity modeled by a gamma distribution with four categories, as selected by ProtTest, was used. In the Bayesian analysis, a relaxed molecular clock approach with log-normally distributed rate was applied. The trees are drawn to the same scale of average amino acid substitutions per site, as indicated by the bar in the middle. Note the considerably shorter branch lengths of the NJ tree compared to the other two trees. Robinson-Foulds distances measuring the topological differences between tree pairs are shown in gray.

After a tree is chosen, it is common to assign support values to internal nodes through assessing the nodes' persistence in trees related to the chosen tree. One particular technique, called bootstrap analysis, in which trees are generated for numerous randomly modified derivatives of the original data set, is most frequently used in distance-based as well as MP and ML analysis. Each internal node in the original tree is characterized by a so-called bootstrap value that is equal to the number of times a node appears in all tested trees. Although the relationship between bootstrap and statistical values is not linear, nodes with very high bootstrap values are considered to be reliable. In a Bayesian analysis, the support of internal nodes is quantified through posterior probability values.

If species evolve according to a molecular clock model, the root position in a tree could directly be calculated from the observed inter-species differences as a midpoint of cumulative inter-species differences. Alternatively, the root position may be assigned to a tree from knowledge about the analyzed species that was gained independently from phylogenetic analysis. Commonly, this knowledge comes in the form of a single or more species which are assumed (or known) to have emerged before the 'birth' of the analyzed cluster. These early diverged species are collectively defined as 'outgroup', while the analyzed species may be called 'in group' (Fig. 3B, C). Also, a tree may be generated unrooted, a common practice in phylogenetic analysis of viruses for which the applicability of the molecular clock model remains largely untested and reliable outgroups may not be routinely available (Fig. 4). In an unrooted tree, grouping of species in separate clusters may be apparent, although these clusters may not be treated as monophyletic as long as the direction of evolution has not been defined. These challenges are addressed by the development of new approaches that infer rooted trees without artificially restricting species evolution to a constant rate (known as relaxed molecular clock models).

Virus phylogeny can be inferred using either genomes or distinct genes and each of these approaches, standard in phylogenomics, may be considered as complementary. Under the first approach, genome-wide alignments are used for analysis. Due to complexities of the evolutionary process that may be region specific, reliable genome-wide alignments can routinely be built only for relatively closely related viruses whose analysis, however, may be further complicated by recombination events (see below). Using the second approach, genes with no evidence for recombination may be merged (concatenated) in a single data set that may be used to produce a superior phylogenetic signal compared to those generated for distinct genes or entire genomes. For viruses with small genomes or for a diverse set of viruses, it is common practice to use a single gene to infer virus phylogeny. Although the results produced may be the best models describing evolutionary history of a group of viruses, the validity of this gene-based approach for the genome-wide extrapolation remains a point of debate. Recently, network methods were used to infer and depict evolutionary relationships of multigene virus genomes taking into account gene-specific sequence affinities.

When the gene tree is used as representing the phylogeny of the entire genome, an underlying most common assumption is that its topology but not branch lengths holds for different genomic regions in reflection of their coevolution with potentially different rates of substitution. This assumption may be violated due to several evolutionary processes, including orthologous gene exchange between (closely) related viruses, gene duplication and horizontal gene transfer (HGT), all involving one or another form of recombination, or incomplete lineage sorting. In phylogenetic terms, this violation may be revealed through incongruency of trees built for different genome regions (Fig. 3F). Trees may also become incongruent due to various technical reasons related to the size and diversity of a virus data set. These characteristics complicate interpretation of the congruency test, which is widely used in different programs to identify recombination in viruses. Other pitfalls of phylogenetic reconstruction include the inability to resolve basal branching patterns of highly divergent lineages (Fig. 3D) and the relatively close clustering of lineages that are only distantly related and do not form a monophyletic group in the true (unknown) phylogeny (Fig. 3E). The latter phenomenon is known as long branch attraction (LBA) and the phylogenetic artifacts produced by LBA are most frequently observed for isolated, that is, long branches in the tree which represent distant lineages with no close relatives known.

Applications of Phylogeny in Virology

Phylogenetic analysis is used in a wide range of studies to address both applied and fundamental issues of virus research, including epidemiology, diagnostics, forensic studies, phylogeography, origin, evolution, and taxonomy of viruses. The first questions to be answered during an outbreak of a virus epidemic concern the virus identity and origin. Answers to these questions form the basis for implementing immediate practical measures and prospective planning, enabling specific and rapid virus detection and epidemic containment, which may include the use and development of antiviral drugs and vaccines. Among different analyses performed for virus identification at the early stage of a virus epidemic, the phylogenetic characterization is used for determining the relationship of a newly identified virus with all other previously characterized and sequenced viruses.

Results of this analysis may be sufficient to provide answers to the questions posed, as regularly happens with closely monitored viruses that include most human viruses of high social impact, for example, influenza virus, human immunodeficiency virus (HIV), hepatitis C virus (HCV), poliovirus, and others. For these viruses, there exist large databases of previously characterized isolates and strains that comprehensively cover the so far characterized natural diversity. Should a newly identified virus belong to one of these species, chances are that it has evolved from a previously sampled isolate or a close variant and this immediately becomes evident in the clustering of these viruses in the phylogenetic tree. Combining the results of gene-specific and genome-wide phylogenetic analysis allows one to determine whether recombination contributed to the isolate origin. For instance, recombination was found to be extremely uncommon in the evolution of HCV, but not for poliovirus lineages that recombine promiscuously, also with closely related human coxsackie A viruses, both of which belong to the same virus species of human enteroviruses known as *Enterovirus C*.

When an emerging infection is caused by a new never-before-detected virus, the phylogenetic analysis is instrumental for classification of this virus and in the case of a zoonotic infection, for determining the dynamic of virus introduction into the (human) population and initiating the search for the natural virus reservoir. This was the case with many emerging infections including those caused by Nipah virus, a paramyxovirus, SARS coronavirus (SARS-CoV), MERS coronavirus (MERS-CoV), ebolavirus, and Zika virus. In the case of SARS-CoV, poor sampling of the coronavirus diversity in the lineage at the time, some uncertainty over the relationship between phylogeny and taxonomy of coronaviruses, and the complexity of phylogenetic analysis of a virus data set including isolated distant lineages led to considerable controversy over the exact evolutionary position of SARS-CoV among coronaviruses. Since then, the matter has fully been resolved but this experience illustrates some challenges in inferring virus phylogeny.

The search for a zoonotic reservoir of an emerging virus may involve a significant and time-consuming effort that requires numerous phylogenetic analyses of ever-expanding sampling of the virus diversity generated in pursuit of the goal. In this quest, phylogenetic analysis canalizes the effort and provides crucial information for reconstructing parameters of major evolutionary events that promoted the virus origin and spread. For instance, intertwining HIV and simian immunodeficiency virus (SIV) lineages in the primate lentivirus tree led to the postulation that the existing diversity of HIV in the human population originated from several ancestral viruses independently introduced from primates over a number of years. Similar phylogenetic reasoning was used to trace the origin of a local HIV outbreak to a common source of HIV introduction through dental practice (known as 'HIV dentist' case). These are typical examples illustrating the utility of phylogenetic analysis for epidemiological and forensic studies.

Geographic distribution of places of virus isolation is another important characteristic relative to which virus phylogeny may be evaluated. This field of study belongs to phylogeography. The evolution of human JC polyomavirus provides an example of confinement of circulation of virus clusters to geographically isolated areas, represented by three continents. Identification of West Nile virus in the USA illustrates a geographical expansion of an Old World virus into the New World. Analysis of phylogenies of field isolates of rabies virus of the family *Rhabdoviridae* sampled from different animals across Europe led to the recognition that interspecies virus expansion occurs faster when compared to geographical expansion.

Phylogenies can also reveal information about the relative strength of the virus–host association over time. In some virus families (e.g., the *Coronaviridae*) host-jumping events may be relatively frequent in establishing new species, including the emergence of at least three human viruses, dead-end SARS-CoV and MERS-CoV and successfully circulating human coronavirus OC43 (HCoV-OC43). At the other end of the spectrum one finds the family *Herpesviridae*. Extensive phylogenetic analysis of herpesviruses and their hosts showed a remarkable congruency of topologies of trees indicating that this virus family may have emerged some 400 million years ago and that herpesviruses largely cospeciate with their hosts. Moreover, through phylogenetic analysis one can show that most viruses, and in particular RNA viruses, evolve at rates that are orders of magnitude faster than those of cellular organisms. For instance, even the most conserved enzymes encoded by nidoviruses, that comprise just four RNA virus families, accumulated more than twice as many substitutions during evolution than their counterparts across the ToL, as estimated through branch lengths of the respective phylogenetic trees (Fig. 1). Taking into account that the MRCA of all cellular organisms predates that of nidoviruses, this reveals that most residues of viral proteins changed repeatedly and more frequently than cellular protein residues during long-term evolution. In fact, this high evolutionary rate seems to be a prerequisite for RNA viruses to stay fit in the ever-changing environment considering their tiny genomes that would otherwise not be able to produce enough genetic variation.

Phylogenetic analysis becomes increasingly important in virus classification (taxonomy) whose development relies on complex multicharacter rules applied to separate virus families by respective 'study groups'. For viruses united in high-rank taxa above the genus level, phylogenetic clustering for most conserved replicative genes is commonly observed and used in the decision making process. For instance, human hepatitis E virus, originally classified as a calicivirus using largely virion properties, was eventually expelled from the family due to poor fit of genome characteristics, including results of phylogenetic analysis. Phylogenetic considerations also played an important role in establishing new families, for example, the *Marnaviridae* and *Dicistroviridae*. In contrast, phylogenetic analysis has been of relatively little use in the taxonomy of large DNA phages which has been developed in such a way that existing families may unite phages with different gene layouts and phylogenies. The relationship between phylogeny and taxonomy is evolving and efforts were made in extracting taxa structure from monophyletic clusters in trees using analysis of pairwise evolutionary distances. In future one might hope for important advancements of virus taxonomy that improve cross-family consistency in relation to phylogeny.

Acknowledgements

AEG research was partially supported by Leiden University Fund and EU Horizon2020 project EVAg 653316.

Further Reading

Dolja VV and Koonin EV (eds.) (2006) Comparative genomics and evolution of complex viruses. *Virus Research* 117: 1–184.

Domingo E (2007) Virus evolution. In: Knipe DM, Howley PM, and Griffin DE, et al. (eds.) *Fields virology*, 5th edn., pp. 389–421, Philadelphia, PA: Wolters Kluwer, Lippincott Williams and Wilkins.

Domingo E, Webster RG, and Holland JJ (eds.) (1999) *Origin and evolution of viruses*. San Diego: Academic Press.

Drummond AJ, Suchard MA, Xie D, and Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

Felsenstein J (2004) *Inferring phylogenies*. Sunderland, MA: Sinauer Associates, Inc.

Gibbs AJ, Calisher CH, and Garcia-Arenal F (1995) *Molecular basis of virus evolution*. Cambridge: Cambridge University Press.

King AMQ, Adams MJ, Carstens EB, and Lefkowitz EJ (eds.) (2012) *Virus taxonomy: the 9th report of the international committee on taxonomy of viruses*. San Diego, CA: Elsevier, Academic Press.

Lauber C, Goeman J, Parquet MDC, Nga PT, Snijder EJ, Morita K, and Gorbalenya AE (2013) The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathogens* 9(7): e1003500.

Moya A, Holmes EC, and Gonzalez-Candelas F (2004) The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology* 2: 279–288.

Page RD and Holmes EC (1998) *Molecular evolution. A phylogenetic approach*. Boston: Blackwell Publishing.

Salemi M and Vandamme AM (eds.) (2003) *The phylogenetic handbook. A practical approach to DNA and protein phylogeny*. Cambridge: Cambridge University Press.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Villarreal LP (2005) *Viruses and evolution of life*. Washington, DC: ASM Press.

Weaver SC, Denison M, Roosinck M, and Vignuzzi M (eds.) (2016) *Virus Evolution, Current Research and Future Directions*. Caister: Academic Press.